ARACNE

MEMORIA

Un proyecto financiado por BBVA

Diciembre 2014-Diciembre 2015

Contenido

Motivaciones y objetivos

La lengua cambia cada día: nacen palabras, mueren otras, algunas adquieren nuevos significados o saltan fronteras. Nacer, morir, transformarse o reconvertirse es parte de la evolución natural del idioma.

La lengua de la prensa es particularmente sensible a estos cambios: azuzado por la inmediatez de la noticia y por la novedad constante, el lenguaje periodístico es un testigo excepcional de los cambios que se producen continuamente en la lengua. En Fundéu BBVA conocemos bien esa transformación constante en la que vive inmerso el idioma. Resolver las dudas lingüísticas que suscita la realidad cambiante de quienes trabajan en los medios de comunicación está en nuestra razón de ser y es a lo que dedicamos nuestro esfuerzo y nuestro tiempo.

Quizá por vivir pegados a la noticia y percibir muy de cerca cómo la lengua cambia día a día, nos resulta inevitable preguntarnos cómo ha evolucionado el lenguaje de la prensa en las últimas décadas. ¿Es hoy más rica la lengua de la prensa que antes o se ha ido empobreciendo, como tantas veces se asegura? ¿Qué palabras se usaban antes más que ahora? ¿Qué palabras han llegado? ¿Cómo influye el transcurso del tiempo en la evolución del lenguaje periodístico?

El interés por la evolución de la lengua no es un tema nuevo en absoluto, y en concreto el debate en torno a la riqueza lingüística suele verse con frecuencia tanto en los medios de comunicación como en las conversaciones diarias del hablante de a pie. ¿Cuántas veces hemos oído decir que el español se está empobreciendo; que cada vez usamos menos palabras; que en el diccionario de la RAE se recogen 88 000 palabras, Cervantes usó unas 8000 en el *Quijote* y, sin embargo, el español medio utiliza unas 300 al día? Se dice que cada vez

se usan menos palabras diferentes en el día a día debido a un desconocimiento cada vez mayor de la diversidad léxica que ofrece nuestro idioma. Es evidente que esta es una preocupación generalizada entre los expertos en lengua española. No obstante, parece que estas afirmaciones no tienen más fundamento que la intuición: hasta ahora no ha habido ningún estudio científico sobre la variación de la diversidad léxica en español que pueda confirmar o refutar esta creencia.

Este es el motor que nos ha movido a Fundéu BBVA y Molino de Ideas a crear el proyecto Aracne. Vamos a dejar a un lado las creencias y las suposiciones infundadas sobre la lengua para observar, con toda la distancia y la honestidad científica a nuestro alcance, cómo ha evolucionado la lengua de la prensa escrita de España desde 1914 hasta 2014. Tomar como objeto de estudio el lenguaje periodístico escrito tiene varias ventajas: por un lado, es relativamente fácil de procesar por ordenadores; en él confluyen registros cultos (artículos de opinión, análisis), coloquiales (deportes, humor) y los más puramente informativos (noticias, reportajes); por último, tiene una gran influencia en la sociedad, tanto en su capacidad para extender y popularizar palabras como para reflejar (y hasta moldear o normalizar) formas de pensamiento.

Analizar los periódicos del siglo XX, por tanto, no solo nos permite estudiar cómo ha cambiado la lengua en las últimas décadas, sino que además nos brinda la emocionante oportunidad de asomarnos a la sociedad y al pensamiento de nuestra sociedad y contemplar los avatares y las transformaciones que han ocurrido en España en los últimos cien años.

Emprendemos un apasionante viaje para descubrir cómo ha evolucionado el lenguaje de la prensa y observar qué nos dicen esos cambios sobre la transformación de la sociedad española.

Consideraciones teóricas

Con el proyecto Aracne pretendemos observar y medir cómo ha cambiado el lenguaje de la prensa española en los últimos cien años, analizando con especial atención los rasgos de riqueza. Pero ¿cómo se mide la riqueza? ¿Qué manifestaciones objetivas y medibles podemos estudiar en los textos para evaluar la riqueza? En definitiva, para medir la riqueza lingüística y su variación a lo largo de los años necesitamos definir qué vamos a considerar riqueza lingüística, y que esa definición se fundamente no sobre apreciaciones subjetivas o consideraciones personales, sino sobre rasgos objetivos, imparciales, mensurables y, sobre todo, comparables.

La variable TTR

Tradicionalmente, en los estudios sobre riqueza lingüística se ha tomado como parámetro primordial la variación léxica, es decir, la cantidad de palabras diferentes que contiene un texto. Esta medición se ha expresado tradicionalmente como la relación entre el número de palabras diferentes que contiene un texto dividido entre las palabras totales de ese texto. Esta relación se llama, en la literatura especializada, *type-token ratio* (TTR), siendo *types* el repertorio de palabras distintas y *token* el número de palabras totales.

$$TTR=Types/Tokens$$

Por ejemplo, una oración como

Frío rigurosísimo, con cielo casi completamente despejado.

tendría un TTR de 1, ya que las siete palabras de esta oración son diferentes (ninguna se repite), por lo que 7/7=1. Mientras tanto, siguiendo la definición tradicional del TTR, una oración como

El comisario señor Flores, con el inspector señor Homar,
ordenó fuera acompañado a la Casa de Socorro.

tendría un TTR de 0.88, ya que de las diecisiete palabras totales de la oración, quince son diferentes (*el* y *señor* aparecen dos veces).

Los valores posibles del TTR oscilan necesariamente entre 0 y 1: sea cual sea el número total de palabras del texto (es decir, sea cual sea el valor que tome la variable *tokens* del denominador y que podemos llamar *n*), el número de palabras distintas de un texto será como mínimo 1 (habrá al menos una palabra distinta en el texto) y como máximo será *n*, es decir, será igual al valor de palabras totales (como hemos visto en el primer caso). Por lo tanto, como mínimo el valor del parámetro TTR será 1/n y como máximo será n/n (es decir, 1).

Las limitaciones del TTR

El TTR cuantifica de una manera simple la noción de riqueza como uso de palabras distintas dentro de un texto. Como medición de la riqueza léxica, resulta bastante intuitivo y simple de calcular. No obstante, presenta tres inconvenientes importantes que sesgan o limitan la observación y a los que hemos tenido que enfrentarnos en el proyecto Aracne para corregir o compensar.

El primer inconveniente del TTR es que el resultado está influido por la longitud del texto que se mide. En una frase muy corta es muy posible que todas las palabras sean distintas, lo que producirá valores de TTR próximos a 1. Sin embargo, según vaya aumentado la extensión del texto, lo esperable es que haya palabras que se repitan. Cuanto más largo se va haciendo un texto, menos probable será que surjan palabras que no hayan aparecido antes. El

grueso de palabras distintas de un texto suele manifestarse en las primeras líneas. Este hecho se ve más claro si pensamos que buena parte de los textos están constituidos por artículos, preposiciones, conjunciones y otras palabras que forman parte de conjuntos cerrados sin significado pleno y con una función más gramatical que semántica. Estas palabras son ineludibles en la redacción de los textos (son los elementos que nos permiten dotar de coherencia gramatical a la expresión) y se repiten mucho, lo que disminuye notablemente el valor de la variable *types* mientras aumenta el de *tokens* al calcular el TTR. Por lo tanto, los textos de mayor longitud tenderán a verse injustamente penalizados con TTR menores, simplemente por el hecho de ser más largos, no por ser verdaderamente menos ricos.

El segundo inconveniente del TTR es que considera como palabras distintas aquellas formas que sean diferentes. Esto se debe a que el TTR nace como medida para caracterizar textos en inglés, que es un idioma con una variación morfológica limitada. Pensemos, por ejemplo, en el artículo determinado *the,* que en español consta de cuatro formas (*el, la, los, las*) cuando en inglés solo tiene una. Un adjetivo tan sencillo y ubicuo como *good* es invariable en inglés, pero en español lo podemos encontrar como *bueno, buena, buenos* o *buenas*, según lo exija la concordancia gramatical. Siguiendo estrictamente la manera de contar que propone el TTR, estas cuatro formas serían palabras distintas. Esta lógica, sin embargo, no parece muy adecuada para una lengua flexiva como el castellano.

Por último, el TTR es una aproximación útil para cuantificar uno de los rasgos habitualmente percibidos como riqueza: el de que mayor diversidad léxica conlleva mayor riqueza lingüística. Sin embargo, la diversidad léxica parece ser solo uno de los aspectos que determinan la riqueza lingüística de un texto. En los últimos años, a la diversidad léxica se le han empezado a sumar otras

características textuales que pueden ayudarnos a evaluar la riqueza lingüística de una manera más completa.

- La sofisticación lingüística, por ejemplo, evalúa el grado de complejidad de un texto a partir del nivel de dificultad del vocabulario empleado o la elaboración de la sintaxis.
- La densidad léxica mide la relación numérica entre el número de palabras de categoría semánticamente plena (sustantivos, adjetivos, verbos) frente al número total de palabras de un texto.
- La existencia de erratas.

La propuesta de Aracne

Vistas estas consideraciones teóricas, nuestro planteamiento a la hora de abordar el proyecto Aracne ha tenido en cuenta los siguientes puntos:

1. Las mediciones de riqueza (TTR o similares) deben hacerse teniendo en cuenta la longitud de los textos medidos para que sean comparables.
2. Conocer la categoría de las palabras es un factor importante para medir la riqueza: las palabras sin carga semántica son, en parte, causantes de la penalización en el TTR que sufren los textos largos; por tanto, resultará interesante poder hacer mediciones que distingan la categoría gramatical de las palabras. Además, podemos asumir que la carga semántica del texto recae sobre sustantivos, adjetivos y verbos, por lo que el conocer la categoría gramatical de las palabras puede sernos de gran utilidad.
3. Dadas las particularidades morfológicas del español, podremos afinar los resultados y obtener mejores cálculos del TTR si agrupamos las diversas formas de una misma palabra bajo su lema, es decir, que *bueno, buena, buenos, buenas* sean considerados en nuestros cálculos como cuatro formas (cuatro *tokens*) de una única palabra (un *type*). De

este modo, nuestras mediciones del TTR tendrán en cuenta el análisis gramatical de la palabra y su lematización.

4. Además de calcular los índices del TTR (con las matizaciones y adaptaciones que se derivan de lo que acabamos de ver), nuestro estudio sobre la riqueza no se limitará solo a la evaluación de la diversidad léxica, sino que también tendrá en cuenta otros aspectos como son la densidad léxica y la complejidad.

Constitución del corpus del proyecto Aracne

Con Aracne queremos observar cómo ha evolucionado el lenguaje de la prensa desde 1914 hasta 2014. Pero ¿por dónde empezar? ¿Qué textos habrá que considerar? El primer paso del proyecto consiste en confeccionar un corpus. Un corpus no es más que una colección de textos reales más o menos numerosa. En nuestro caso, necesitaremos confeccionar un corpus que cubra el intervalo temporal que queremos analizar (1914-2014) para que sirva de muestra de la lengua periodística de la época. Necesitaremos que, en conjunto, nuestro corpus resulte lo más representativo y equilibrado posible para que el análisis sea fiable.

La selección del corpus del proyecto Aracne nos ha suscitado muchas cuestiones que, además de suponer un reto metodológico para nuestro proyecto, nos han dado pie a reflexionar sobre la naturaleza del lenguaje y la evolución del idioma.

Selección de periódicos

La primera cuestión ineludible es acotar qué características deben cumplir los periódicos candidatos para ser incluidos en nuestra selección. No podemos analizar exclusivamente ejemplares de un solo periódico porque las mediciones entonces resultarían sesgadas: no sabríamos si nuestras observaciones se deben a características propias del periódico escogido o si son verdaderamente generalizables al conjunto del lenguaje periodístico de una época. Por lo tanto, nuestro corpus deberá estar conformado por ejemplares de distintos periódicos.

Otra condición irrenunciable es que las fuentes han de tener orígenes diversos. La variación geográfica es uno de los rasgos diferenciadores de la lengua. Si bien este hecho es más acusado en la oralidad, no debemos perderlo de vista y

hemos de considerar periódicos procedentes de distintos puntos de la península para tener más pluralidad y variedad lingüística regional.

Los periódicos escogidos han de ser generalistas para que la variedad léxica sea lo más completa posible. Si introducimos en el estudio periódicos de temática muy concreta (economía, deportes...), estaremos comparando ejemplares de vocabulario muy específico con ejemplares generalistas (previsiblemente de vocabulario más variado), lo que podría acarrear sesgos en la medición de variaciones en la riqueza.

Por último, nos encontramos con una restricción insalvable puramente material: solo podremos considerar para nuestro estudio aquellos periódicos que cuenten con ejemplares digitalizados disponibles. Esta condición nos limita enormemente la selección, pero resulta inevitable, puesto que el procesamiento del texto será automático. No es difícil dar con hemerotecas digitalizadas de periódicos posteriores a los años 70. Sin embargo, acceder a digitalizaciones que cubran el intervalo de años entre 1914 y 1970 resulta mucho más complicado.

Con estas condiciones de partida, hemos procedido a seleccionar los periódicos que tendríamos en cuenta para el estudio. Queremos agradecer a *El Norte de Castilla, El Correo, Las Provincias,* al *Diario de Mallorca, Diario La Rioja, Heraldo de Aragón, ABC* y a la Biblioteca Nacional su colaboración imprescindible en esta parte del proceso. Finalmente, por motivos de accesibilidad, estado de la digitalización, disponibilidad de ejemplares, cobertura temporal y equilibrio del corpus han sido seleccionados ejemplares de *El Norte de Castilla, El Correo de Mallorca, La Almudaina,* del *Diario de Mallorca* (fruto de la fusión de los dos anteriores), del *Heraldo de Aragón* y de *La Vanguardia*, repartidos homogéneamente a lo largo del tiempo.

Que la selección de periódicos se mantenga constante nos asegura una homogeneidad en la composición muy valiosa de cara al análisis de los datos. No obstante, esta decisión conlleva una desviación ideológica que, sin saber

con certeza si causa sesgos en el estudio, no queremos dejar de mencionar: necesariamente, los periódicos que han perdurado desde 1914 hasta 2014 son aquellos que sobrevivieron al franquismo y, en consecuencia, conllevan una inclinación ideológica difícil de obviar. No perdamos de vista el objetivo final del estudio: la medición de la riqueza léxica. Si bien parece legítimo asumir que la ideología puede condicionar el vocabulario de un texto (usando unas palabras, eliminando otras), no sabemos hasta qué punto es posible achacar a motivos ideológicos diferencias en la riqueza (esto es, en la variación y la sofisticación de un texto, no en el repertorio del léxico). Consideramos importante mencionar esta cuestión, no solo ya como parte de la descripción del corpus de Aracne, sino también con la intención de lanzar un guante para quien quiera recogerlo e investigar sobre la relación entre ideología y riqueza léxica.

Selección de ejemplares

Una vez que ya tenemos la selección de periódicos que van a formar parte del estudio, es necesario decidir qué ejemplares vamos a incorporar al corpus. ¿Cuántos ejemplares cogeremos? ¿Repartidos de qué manera a lo largo del tiempo?

Lo deseable hubiera sido incorporar al estudio las hemerotecas completas de los periódicos seleccionados. Lamentablemente, la digitalización de los periódicos se encuentra en formato imagen, no en formato textual. Para poder analizar la lengua de los periódicos necesitamos disponer de los textos de los ejemplares, lo que conlleva necesariamente el procesamiento de las imágenes escaneadas mediante un sistema de reconocimiento óptico de caracteres (OCR, Optical Character Recognization). La tecnología del OCR dista mucho de ser perfecta. Sin desmerecer la inestimable ayuda que nos ofrece, es habitual encontrar errores de reconocimiento en los textos producidos por el OCR, sobre todo en ejemplares antiguos, donde la calidad de la imagen es

más deficiente. Nuestro afán por observar de manera rigurosa y pormenorizada la variación léxica de la prensa no puede permitirse sustentar un estudio sobre una inmensa cantidad de textos defectuosos. Por lo tanto, nos vemos obligados a que los textos producidos por el OCR sean supervisados por un revisor humano que garantice una calidad aceptable antes de ser incorporados al estudio. Por consiguiente, la cantidad de ejemplares para seleccionar se ve acotada por una nueva restricción: debemos limitarnos a un número que pueda ser revisado manualmente en un tiempo razonable y ajustándonos a los recursos humanos de los que dispone el estudio.

Puesto que no podemos procesar las hemerotecas completas de los periódicos seleccionados, sino solo una parte de ellas, necesitamos decidir qué ejemplares vamos a incluir. Una variable fundamental del estudio es el tiempo. A fin de cuentas, lo que pretendemos medir es cómo ha cambiado la riqueza a lo largo de unos años concretos. Surge entonces la cuestión de cómo debemos distribuir los ejemplares que seleccionemos de ese período para poder medir variaciones. Es decir, si lo que queremos medir es cómo ha cambiado la lengua a lo largo del tiempo, tendremos que tomar fotografías lingüísticas de la prensa en distintos momentos del intervalo de años que queremos estudiar y compararlas después. Esto entraña varias preguntas de difícil solución: ¿cuándo tomar esas fotografías?, ¿cada cuántos años tomar muestras? y ¿cuál es la unidad de tiempo en lo que a cambio lingüístico se refiere?

Ante esta cuestión, hemos barajado dos posibilidades. Una opción consiste en seleccionar unas pocas muestras muy concentradas en unos años muy concretos, estudiar la lengua de esas muestras y asumir que las observaciones hechas para esos años serán extrapolables a otros ejemplares de la época; por ejemplo, podríamos hacer tres muestras pormenorizadas, una de 1914, otra de 1964 y otra de 2014, y estudiar con detalle la lengua de cada uno de esos años. Siguiendo esta forma de proceder, asumiríamos que las diferencias

encontradas entre las observaciones de la muestra de 1914 y de 2014 se deben a tendencias globales que afectan a la lengua de cada época. Esta modalidad de catas escasas y profundas tiene la ventaja de que caracteriza con gran definición la lengua de un año concreto, pero tiene como contraparte que sesga enormemente la observación. Pensemos, por ejemplo, en la muestra que correspondería al año 1914 siguiendo esta aproximación: la Guerra Mundial fue un tema fundamental en los periódicos de ese año, lo cual desvía el léxico hacia campos semánticos muy concretos. No podemos asumir que el análisis léxico de un año concreto (con toda la desviación temática que eso conlleva) pueda representar fielmente el perfil léxico del período que abarca desde 1914 hasta 1964, año del que se efectuaría la siguiente cata. Esta aproximación, además, conlleva catas muy espaciadas en el tiempo e intervalos de años muy largos sin datos, dando por sentado que las variaciones entre una cata y otra serán homogéneas y graduales. Es decir, si nos encontrásemos unos índices de riqueza muy elevados en 1914 y más bajos hacia 1964, ¿podríamos asumir que la caída se ha producido de forma gradual y escalonada en esos cincuenta años? ¿O quizá los índices de riqueza son más inestables y lo que asumiríamos como variaciones graduales esconden en realidad picos y valles? La historia de la lengua nos enseña, además, que si bien los cambios gramaticales suelen ser lentos y se extienden durante generaciones, las variaciones léxicas están muy relacionadas con los cambios culturales e históricos y son notablemente más rápidos. Por lo tanto, hemos optado por descartar esta primera aproximación.

Desechada esta distribución de catas muy concentradas, muy espaciadas en el tiempo y muy pormenorizadas, hemos optado por una distribución temporal más extendida. Hemos seleccionado ejemplares de todo el período que vamos a estudiar, desde 1914 hasta 2014, para acumular después los datos en intervalos de diez años cuando las mediciones son cualitativas (rasgos léxicos) y de veinte años para las cuantitativas (densidad y variación). Esto quiere decir que los datos que extraigamos de la observación de 1914 irán diluidos en

el agregado de datos que representan la década desde 1914 hasta 1923 en el caso de las mediciones de frecuencia léxica y al intervalo entre 1914 y 1933 en el caso de las mediciones cuantitativas. Esta disolución nos asegura que las referencias temporales, temáticas, históricas o azarosas estarán compensadas con el resto de las observaciones de la década, que además corresponderán a otros años para no sesgar históricamente la observación.

Hemos seleccionado la década como unidad temporal léxica por ser un intervalo lo suficientemente amplio como para poder observar los cambios históricos, culturales y sociales que se reflejan en el lenguaje periodístico, pero suficientemente corto como para que las variaciones que midamos sean progresivas y no enmascaren picos. En cuanto a las mediciones cuantitativas (densidad y variación), la acumulación de datos en intervalos de veinte años es la que permite observar con mayor claridad la variación. Esta aproximación conlleva necesariamente que todas las décadas estén representadas, de una manera más o menos homogénea, para después poder comparar unas con otras. En un primer momento, esto nos llevó a hacer una selección constante de ejemplares por década, pero nos encontramos con una nueva disyuntiva: los periódicos de principio de siglo son significativamente más cortos y se van alargando con el paso de las décadas. Esto quiere decir que si lo que mantenemos constante es el número de ejemplares por década, tendremos algunas décadas representadas con más artículos y más palabras que otras. Así que hemos optado por mantener homogéneo el número de palabras por época, y, por tanto, el número de ejemplares por década es más alto a principio de siglo y disminuye a medida que los periódicos van siendo más extensos.

Con estas cuestiones resueltas, solo queda decidir qué fechas concretas serán las seleccionadas. Esto plantea otras preguntas interesantes que aquí esbozamos y sobre las que consideramos que merecería la pena profundizar: teniendo como principal objetivo la comparabilidad entre épocas, ¿es preferible fijar unas fechas concretas en las que seleccionar los ejemplares

para minimizar la variación estacional que pudiera existir? Es decir, si lo que queremos es ver cómo varía la riqueza a lo largo del tiempo, tendremos que intentar minimizar los cambios que puedan venir causados por otras variables, como, por ejemplo, la época del año. ¿Es posible que la riqueza léxica varíe con los meses y, por lo tanto, sea tramposo comparar la riqueza de ejemplares de febrero con los de septiembre? ¿Sería interesante seleccionar una fecha anodina sobre la que tomemos todos los ejemplares a lo largo de los cien años para ver cómo varía la riqueza, con iguales condiciones estacionales? Aunque esta vía es tentadora, una vez más, la restricción de variación en aras de la comparabilidad nos sesgaría el estudio. Pensemos que si seleccionáramos ejemplares de verano tendríamos una sobrerrepresentación de campos léxicos muy concretos (mayor temática de ocio, probablemente menor de vida política) y, consecuentemente, de determinadas palabras. Algo parecido ocurre con la selección del día de la semana. ¿Existen diferencias léxicas relevantes según el día de la semana en que esté publicado el artículo? Aunque a primera vista esta cuestión pueda parecer baladí, tiene su intríngulis: y es que en domingo probablemente haya más artículos de corte editorial (es decir, textos de opinión), mientras que los viernes suelen publicarse reseñas y críticas de espectáculos y ocio.

Por lo tanto, tras estas reflexiones que acabamos de exponer, la selección de ejemplares se ha hecho:

- considerando la década la unidad temporal mínima sobre la que vamos a trabajar, si bien los distintos ejemplares que representen a una década habrán de ser de años distintos;
- primando la homogeneidad en el número de palabras, no en el número de ejemplares;
- seleccionando aleatoriamente las fechas concretas para garantizar variación tanto en los meses como en los días de la semana representados.

Modus operandi

Una vez escogidas las hemerotecas de los periódicos y seleccionados los ejemplares para el estudio, el *modus operandi* del proyecto Aracne ha consistido en:

1. Obtención de los ejemplares seleccionados para el estudio en formato imagen digital.
2. Procesamiento mediante tecnología OCR para extracción del texto del ejemplar.
3. Supervisión humana (asistida por ordenador mediante el uso de macros, corrección semiautomática y expresiones regulares) de los textos producidos por el OCR para garantizar una calidad aceptable y separación del texto continuo del ejemplar producido por el OCR en artículos.
4. Procesamiento lingüístico de los textos mediante tecnología de procesamiento de lenguaje natural (PLN).
5. Medición de los rasgos de variación léxica (TTR), densidad y complejidad de los textos.
6. Cálculo de medias, desviaciones y agrupación de los datos en intervalos de tiempo.
7. Visualización de los datos, análisis y conclusiones.
8. Recopilación de las curiosidades históricas y lingüísticas de los recortes de prensa encontrados, confección de la página web, redacción de la memoria y publicación.

La unidad mínima de la que disponemos al comienzo del procesamiento es el texto producido por el OCR, es decir, la unidad de partida es el ejemplar de periódico como un todo indivisible. Durante la revisión humana de los textos producidos por el OCR, ese todo continuo es fragmentado en artículos independientes. Son esos artículos independientes (convenientemente identificados por fecha y periódico de procedencia) los que entran en el

engranaje de procesamiento lingüístico para ser separados en oraciones (*splitting*), que a su vez serán lematizadas y analizadas morfológicamente. La lematización es la técnica que permite asignar a cada palabra de una frase su lema, es decir, la forma canónica bajo la que nos la encontraríamos en un diccionario. Esto nos permite agrupar correctamente todas las formas conjugadas de un mismo infinitivo o las diversas variaciones de género y número de sustantivos y adjetivos. Gracias a la lematización, en el proyecto podremos contabilizar *voy*, *iremos* o *hubierais ido* como apariciones del verbo *ir*, o tanto *españoles* y *españolas* como formas de *español*. La lematización aplicada, además, tiene en cuenta el contexto sintáctico de la palabra, lo que permite desambiguar categorialmente las palabras homónimas, como *meses*, que será apropiadamente etiquetada como sustantivo en una oración como *los meses del año* y como verbo en *que te meses las barbas*. La lematización ha sido llevada a cabo con la tecnología lingüística de Molino de Ideas, y los resultados obtenidos de dicha lematización nos han servido para poder hacer los cálculos de variación léxica y densidad.

Hemos calculado dos índices para la variación léxica, ambos inspirados en la relación *types/tokens* (TTR), pero ligeramente modificados para adaptarlos mejor al propósito del proyecto Aracne. La primera cuestión fundamental que vimos al hablar de las limitaciones teóricas del TTR es que no debemos, bajo ningún concepto, comparar valores de TTR entre textos de distinta extensión. Por lo tanto, las mediciones y comparaciones de nuestros índices TTR (o sus modificaciones) se han hecho sobre textos de extensión semejante. Por otro lado, también discutimos en las consideraciones teóricas las limitaciones de partida que tiene el TTR al aplicarlo a lenguas flexivas como el español, ya que la medición tradicional del TTR considera como tipos distintos (*types*) formas distintas (plurales, femeninos, conjugaciones) de una misma palabra. Nuestro cálculo del TTR, por consiguiente, lo hemos hecho considerando las formas lematizadas de las palabras, no la diferenciación tradicional de *types* y *tokens*. Es decir, hemos considerado como *types* el número de lemas distintos

de un texto y como *tokens,* el número de palabras totales. Mientras que el TTR tradicional considera que en «Donde dije digo, digo Diego», *digo* y *dije* son dos tipos independientes (uno con una aparición, el otro con dos), nosotros consideramos que ambas son formas del lema *decir*; así pues, en esta frase nuestro cálculo del TTR computaría un único lema distinto (*decir*) para las tres apariciones (*digo*, *digo* y *dije*). Consideramos que esta aproximación es más apropiada que la convencional porque recoge de una manera más fiel el funcionamiento morfológico del español y lo que se considera variación léxica.

El segundo índice TTR que hemos calculado consiste en una variante aún más restrictiva de nuestra propuesta de TTR. En este segundo índice, el TTR se ha calculado incluyendo en el cómputo (tanto de *types* como de *tokens*, o, en nuestro caso, tanto de formas como de lemas) solo aquellas palabras consideradas semánticamente plenas, es decir, sustantivos, adjetivos, verbos y adverbios terminados en *-mente*. Puesto que las palabras semánticamente vacías (preposiciones, conjunciones, artículos...) no añaden variaciones a la riqueza léxica y engordan artificialmente el recuento de formas (son palabras ineludibles para construir un discurso coherente, pero la mayoría de ellas aparecerán al comienzo del texto y después solo se repetirán) y dado que disponemos de la categoría morfológica de las palabras gracias al procesamiento lingüístico previo, hemos optado por hacer este segundo cálculo solo con las palabras que verdaderamente aportan variación léxica. La comparación de los resultados de medir esta variante semántica del TTR se ha hecho también sobre textos de longitud semejante.

Tanto el cálculo del TTR lematizado como del TTR semántico se han hecho tomando el artículo como unidad de análisis. En ambos casos, los valores del TTR oscilan entre 0 y 1, aproximándose a 1 cuanto mayor es la variación (más igualada está la variación entre lemas diferentes y palabras totales) o a 0 según resulta más repetitivo léxicamente el texto.

El proceso de lematización y categorización morfológica es lo que también nos ha permitido calcular la segunda variable relacionada con la riqueza: la densidad léxica. Hemos medido la densidad como la relación entre el número de palabras con categoría semántica (nombres, adjetivos, verbos, adverbios acabados en *-mente*) entre palabras totales del texto. Los valores de la densidad léxica también oscilan entre 0 y 1, tendiendo a 1 los textos muy densos (es decir, con una alta proporción de palabras semánticamente plenas) y a 0 aquellos en los que aparecen muchas palabras más gramaticales o estructurales.

Por último, la última variable que hemos medido para determinar la riqueza léxica de los textos ha sido la complejidad. Para ello, hemos recurrido de nuevo a la tecnología lingüística de Molino de Ideas. Un programa informático que evalúa distintos grados de dificultad lingüística ha analizado cada artículo y le ha asignado una puntuación del 0 al 10, siendo 0 los textos más complejos y 10 los más sencillos. Entre los rasgos analizados por este programa están la sofisticación del vocabulario, la longitud de las oraciones, la estructura oracional, los tiempos verbales y el grado de referencialidad y abstracción. Hay que tener en cuenta que esta medición es cuestionable en un aspecto fundamental: la complejidad del vocabulario se tiene en cuenta a partir de lo habitual que es una palabra. Sin embargo, la frecuencia de uso de una palabra está ligada a una época histórica. Es posible que hoy consideremos como muy infrecuentes palabras que eran absolutamente habituales en su momento. Para solventar este problema necesitaríamos contar con unas mediciones de frecuencia para las distintas épocas que cubre el corpus de Aracne. A pesar de esta limitación (que solo afecta a una de las múltiples variables que evalúa el programa), hemos decidido analizar igualmente este rasgo, asumiendo que lo que se considera complejo o infrecuente puede variar con el transcurso del tiempo.

Interpretación de los datos y conclusiones

En el apartado *Visualización de datos* de la web de Aracne están disponibles las gráficas que muestran los resultados obtenidos del proyecto Aracne. Hemos distinguido tres grupos de resultados en función de la cuestión que tratan.

Riqueza

En el apartado *Riqueza* se muestran los datos relativos a las mediciones de riqueza lingüística. Tal y como se mencionaba en el apartado *Modus operandi*, han sido tres las variables analizadas:

- La relación TTR en nuestra particular adaptación lematizada: lemas distintos entre palabras totales (primera gráfica) y lemas distintos con categoría semántica entre palabras totales con categoría semántica (segunda gráfica). Los valores oscilan entre 0 y 1.
- La densidad léxica: palabras con categoría semántica (nombres, adjetivos, verbos, adverbios acabados en -*mente*) entre palabras totales. Los valores oscilan entre 0 y 1.
- La complejidad del texto, calculada como el valor medio ponderado sobre distintos rasgos de sofisticación lingüística del texto (complejidad sintáctica, tiempos y modos verbales, dificultad de las palabras utilizadas, referencialidad y abstracción). Los valores oscilan entre 0 y 10.

Las cuatro gráficas que se derivan de estas mediciones se han representado siguiendo una estructura común. Por un lado, dada la naturaleza de la variable TTR y el sesgo que produce (véase el apartado *Consideraciones teóricas*), la agregación de datos se ha hecho teniendo en cuenta la extensión de los textos. Es decir, la comparación de los valores de riqueza se ha hecho entre artículos de longitud semejante. Se distinguen así siete grupos en función del número de palabras del artículo (indicado en el eje horizontal de la gráfica). Si bien

esta restricción sobre extensión de los textos y comparabilidad es solo propia de la medición del TTR (y sus derivados), hemos optado por mantenerla también en la comparación de la densidad y la complejidad para comprobar si hay diferencias reseñables en los valores obtenidos según la longitud del texto.

Por otro lado, los datos de riqueza se han agrupado en intervalos de veinte años. Es decir, hemos fraccionado el intervalo de años entre 1914 y 2014 en cinco épocas, y la información de riqueza sobre esos cinco bloques son los que hemos agregado. Cada una de las barras verticales representadas en las gráficas corresponde a un intervalo de años. Esta agrupación de años ha sido la que mostraba una representación más homogénea, más comparable y permitía una visualización de datos eficaz. No obstante, los valores en bruto sin agregar están en la web disponibles para descargar, para quien quiera volver sobre ellos o analizarlos individualmente. Sobre las mediciones representadas por las barras verticales se ha trazado una línea que representa el valor medio de la variable analizada sobre el total del corpus de Aracne, es decir, el cómputo de la media global sin distinción de épocas. De este modo, podemos observar si los valores obtenidos para una determinada extensión en una época concreta están por encima o por debajo de la media global.

Las cuatro gráficas obtenidas de la medición de las tres variables referidas a la riqueza revelan datos muy homogéneos. Las dos gráficas relativas al TTR muestran un descenso de los índices de variación a medida que los textos se hacen más largos, siendo más acusado el descenso en la primera gráfica. La espectacularidad de este descenso no debe ni alarmarnos ni desviarnos de nuestro análisis de resultados, puesto que forma parte del resultado esperable. Como comentamos en las consideraciones teóricas sobre la riqueza, los índices de variación léxica tienden a disminuir según aumenta la extensión del texto analizado porque la probabilidad de que surjan palabras nuevas que no hayan aparecido antes disminuye según se alarga el texto. La segunda gráfica (correspondiente a nuestra adaptación del TTR lematizado teniendo en cuenta

categorías semánticas) muestra, por tanto, un descenso menos pronunciado porque se han excluido del cómputo las preposiciones, artículos, conjunciones y otras palabras gramaticales que no aportan variación léxica.

Lo que debemos comparar, en consecuencia, son las diferencias entre las cinco épocas (las cinco barras verticales) en cada uno de los valores de extensión del texto (y no las diferencias de riqueza entre textos de distinta longitud, puesto que ya sabemos que esa comparación está sesgada por la propia naturaleza de la variable TTR). Lo que se observa es que los valores se mantienen en general muy estables, con diferencias mínimas entre épocas. Están además muy concentrados y las diferencias respecto a la media nunca superan el 10 % y se mantienen sorprendentemente uniformes en todas las épocas y para todas las longitudes.

Por otro lado, los valores relativos a la densidad muestran también una constancia reseñable para textos de extensión superior a las cien palabras. La densidad media oscila en torno al 0.5 para todas las épocas en textos de más de cien palabras, con variaciones mínimas en la segmentación por épocas. Es interesante observar lo que ocurre con la densidad para los textos de menos de cien palabras. En este caso, la densidad es notablemente más alta en todas las épocas y muestra un máximo absoluto llamativo para los textos de menos de diez palabras de época reciente (1994-2014): 0.926, cuando el resto de épocas tiene una densidad para esa extensión rondando el 0.67. La mayor densidad léxica que muestran todas las épocas en los textos de menos de cien palabras puede achacarse a la redacción tan particular que tienen los titulares y las entradillas breves (que son los textos periodísticos que encontramos con esta extensión). La naturaleza casi telegráfica en la redacción de breves y titulares puede explicar que resulten tan densos. Es decir, estos microtextos periodísticos caracterizados por la elisión de artículos (*Manifestaciones en toda España*) y de palabras que sean poco relevantes para el titular sobresalen en lo que a proporción de sustantivos, adjetivos, verbos y adverbios acabados

en *-mente* se refiere. Esta tendencia se observa más acusada en el máximo absoluto que muestra la gráfica en época reciente. Podemos aventurar un motivo que justifique esta observación: esta densidad tan alta podría deberse quizá a la proliferación de entradas y titulares sintéticos en la primera plana de los periódicos de nuestro tiempo. Si observamos la primera página de un ejemplar antiguo, veremos que la primera plana ya contiene columnas y artículos completos, mientras que en los últimos veinte años, la primera página de los periódicos se ha convertido casi en un índice de avances informativos en forma de titular y texto mínimo que adelantan lo que se detalla en el interior del periódico ya en forma de noticia desarrollada. Es decir, en los últimos veinte años, se puede haber producido una *telegraficación* de titulares y breves en la prensa que explicaría estos valores de densidad.

En cualquier caso, lo que sí podemos asegurar es que la densidad del texto también está influida por su extensión (algo que ya sabíamos que ocurría en las mediciones del TTR, pero no sabíamos si afectaría también a otros índices), ya que todas las épocas muestran densidades superiores cuando los textos son más cortos, aunque a partir de las cien palabras se estabilizan en una proporción de una palabra semánticamente plena por cada dos palabras totales. Es decir, la proporción entre palabras semánticas y gramaticales no es uniforme, sino que, a la luz de lo que muestran los resultados de Aracne, es más alta en textos más cortos. Como hemos visto, es posible que esta observación esté relacionada con el formato mismo de la prensa y sea, por tanto, propia del lenguaje periodístico, y concretamente achacable a los cambios en la maquetación de los periódicos, por lo que no podemos asegurar que esta observación sea extrapolable al conjunto de la lengua en general. Por ello, consideramos que para poder confirmar nuestras sospechas, sería interesante profundizar en la naturaleza de las mediciones de densidad en otros tipos de textos para poder comprobar si, efectivamente, nos encontramos ante una variable lingüística que cambia según la extensión del texto (sea cual

sea el tipo de texto) o si se ha modificado en concreto en la lengua de la prensa a causa del formato. De ser así, podríamos estudiar qué otros cambios de formato han producido variaciones lingüísticas y si el último gran cambio hacia la digitalización de la prensa sigue esta tendencia.

La gráfica de complejidad confirma la innegable estabilidad que presenta la riqueza lingüística. Los valores para todos los intervalos temporales y en todas las extensiones son muy homogéneos. Consideramos esta gráfica particularmente relevante. Las mediciones de variación léxica y densidad tienen dos limitaciones importantes: por un lado, el sesgo que experimentan en relación con la longitud del texto analizado, y, por otro, que en realidad analizan la noción de riqueza desde una aproximación exclusivamente cuantitativa, cuando, desde nuestra experiencia como hablantes, la riqueza es percibida como una noción fundamentalmente cualitativa, es decir, relacionado con el nivel de sofisticación y elaboración de un texto. Con toda la limitación que la medición de la complejidad conlleva (y que expusimos en el apartado *Modus operandi*), estos valores son los que verdaderamente nos permiten acercarnos a la composición del texto, entendiendo la riqueza como un valor asociado a la naturaleza y calidad del contenido, y no solo a un recuento léxico y categorial que, si bien es interesantísimo e insustituible, resulta parcial. La estabilidad que nos revela la gráfica de la complejidad confirma que, más allá de las consideraciones personales o las impresiones subjetivas, la riqueza y la complejidad de los textos periodísticos se han mantenido notablemente estables en los últimos cien años.

Estilo y léxico

Si bien el objetivo fundamental del proyecto Aracne ha sido el estudio de la evolución de la riqueza, no podemos resistirnos a mostrar otros rasgos fascinantes relativos a la evolución lingüística de la prensa que, aunque no

estén directamente relacionados con la riqueza, han aparecido durante el estudio y están disponibles en la sección *Visualización de datos* de la web.

En el apartado sobre estilo, creemos que merece la pena no perderse la evolución que han sufrido los adjetivos en grado superlativo (-*ísimo*) en el lenguaje de la prensa. El análisis de datos confirmó la extinción que a primera vista en la supervisión manual ya nos llamó la atención. Se observa un uso notable del superlativo en los primeros años que cubre el estudio; uso que se desploma según avanza el tiempo. Es posible que esta extinción sea propia del lenguaje de la prensa, no necesariamente del español en general. El superlativo tiene una connotación que hoy resulta excesivamente subjetiva para la neutralidad del lenguaje periodístico, así que bien esa percepción subjetiva es relativamente reciente, bien es posible que según se fue afianzando el estilo periodístico hacia la neutralidad impersonal más que hacia la crónica, este tipo de adjetivos fueran resultando menos apropiados.

Es digna de mención también la evolución del modo. Aunque es cierto que presenta un máximo absoluto en la primera década del estudio (1914-1923), parece mantenerse con ciertas oscilaciones pero bastante viveza. Habrá que seguir observando la evolución del lenguaje de la prensa para poder comprobar si el subjuntivo está en retroceso, como afirman algunas voces. Si lo está, desde luego sus tiempos de evolución son superiores a los cien años considerados en el proyecto Aracne.

También hemos incluido en esta sección las evoluciones de los tratamientos de persona *don* y *señor* (con una espectacular caída en el transcurso del siglo) y la frecuencia relativa de algunos de los anglicismos (tanto en forma cruda como en forma adaptada) propios del mundo de la prensa y que contaban con suficientes apariciones en la muestra como para poder observar su evolución.

En la sección dedicada al léxico, mostramos las gráficas de evolución de la frecuencia relativa a diferentes términos ligados a los avatares culturales e históricos de los últimos cien años, agrupándolos por temática, y la

combinatoria léxica de las palabras *guerra* y *trabajador* a lo largo del tiempo. Hemos escogido mostrar estas palabras y no otras porque, dada su presencia constante en el corpus de Aracne, permiten hacer un interesante viaje a través de la problemática social e histórica de los últimos cien años. Podemos asomarnos, pues, con estas gráficas al perfil léxico que cada época emana en Aracne. Los picos y valles que dibuja la frecuencia relativa de palabras como *guerra*, *peseta*, *fanega*, *comunismo*, *alemán*, *nuclear* o *europeo* nos invitan a viajar por la historia del último siglo y a atisbar cómo era la sociedad que producía y redactaba estas noticias. No obstante, ante las gráficas de frecuencias relativas y los cuadros de combinatoria léxica, es necesario recordar que el análisis de los campos semánticos no ha sido el objetivo primordial del proyecto Aracne, sino un feliz descubrimiento colateral que no podemos dejar de compartir. Sin embargo, por muy disfrutable que sea este subproducto de Aracne, no debemos olvidar que, puesto que la selección de ejemplares se orientó en todo momento para conseguir una muestra representativa y válida para el estudio de la riqueza, no podemos presuponerle la misma representatividad para mostrar de forma fiable el léxico y la variación en los campos semánticos.

Reflexiones finales e hilos para seguir tejiendo

No conozco a nadie a quien no le interese el lenguaje, dice el psicolingüista Steven Pinker. Y es que la lengua nos rodea, nos construye y es el cristal a través del cual observamos el mundo y nos observamos a nosotros mismos. Las palabras que entran o salen del diccionario, los neologismos que se cuelan en nuestros medios o el último anglicismo de moda dan pie a debates encendidos que vuelven una y otra vez sobre el uso actual que hacemos del idioma. Sin embargo, a pesar de la pasión con la que defendemos nuestras creencias lingüísticas, apenas tenemos información real que nos muestre de manera objetiva cómo usamos la lengua. En estos tiempos en que la disponibilidad de información llega hasta el empacho y la tecnología lingüística nos brinda las mejores herramientas posibles de análisis textual, las posiciones en torno a la lengua siguen siendo más propias de la alquimia que de la ciencia.

El proyecto Aracne aspira a ser el punto de partida que dé pie a estudiar de una manera empírica, científica y rigurosa la lengua en los medios de comunicación. En esta primera fase hemos realizado un estudio limitado pero ambicioso en el que nos hemos concentrado en las mediciones relativas a la evolución de la riqueza del lenguaje de la prensa, por ser un tema que suele suscitar gran debate, pero en el que siempre se echan en falta datos objetivos sobre los que sustentar una opinión fundada.

Tras meses de recopilación de datos y análisis lingüísticos y siempre con la prudencia de que todo estudio está sujeto a revisión en su aproximación, métodos y conclusiones, si nos atenemos a los resultados que ofrece el proyecto Aracne, podemos concluir que la riqueza lingüística en términos generales no parece haber sufrido grandes variaciones en el último siglo. A pesar de la creencia generalizada de que la lengua (y en concreto la de los

medios de comunicación) está empobreciéndose, los datos parecen indicar que las variables que se consideran indicadores de la riqueza lingüística son más estables de lo que se suele suponer. Invitamos, no obstante, a continuar profundizando en esta línea y a analizar la riqueza lingüística en la prensa desde otras perspectivas, tratamientos y marcos de estudio distintos a los que nosotros hemos seguido. En la documentación del proyecto hemos apuntado varias decisiones metodológicas que hemos tenido que tomar en el transcurso del estudio: todas son susceptibles de ser revisadas, cuestionadas y replanteadas, y esperamos ver más proyectos en esta dirección para poder contrastar, ampliar y matizar los resultados de nuestra investigación.

Durante el proyecto han ido surgiendo con frecuencia hilos de investigación de los que tirar. Hemos resistido la tentación de desviarnos por caminos que nos alejaban de nuestro objetivo de partida (la riqueza lingüística), pero consideramos imprescindible retomar en el futuro próximo algunas de estas cuestiones para estudiarlas como se merecen. El estudio de los campos semánticos a lo largo del tiempo, la representación del léxico, la evolución de los perfiles lingüísticos, la variación gramatical y la caracterización temporal, ideológica y regional de los medios de comunicación son asuntos fascinantes que piden líneas de investigación propias y constituyen el camino natural para continuar la senda que hemos abierto con el proyecto Aracne. Las gráficas de la frecuencia relativa y la combinatoria léxica son un aperitivo muy apetecible de lo que podemos llegar a analizar.

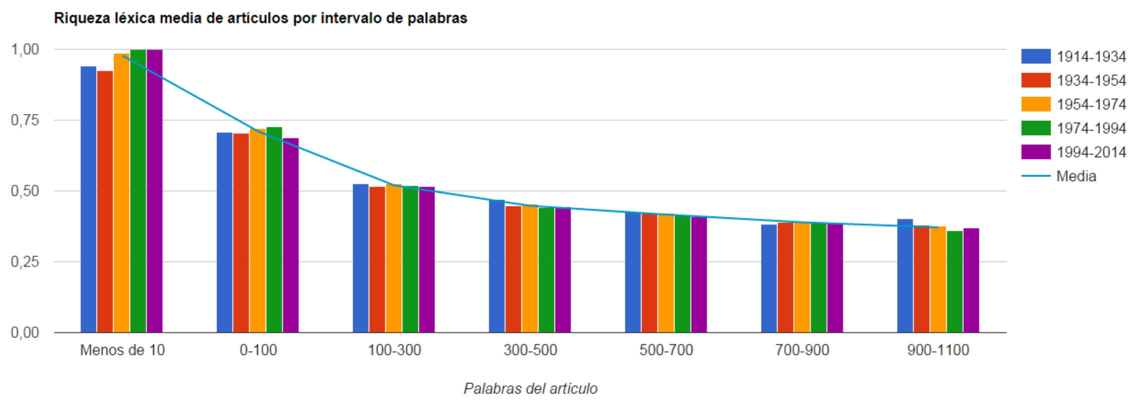
La variedad social, geográfica e histórica que nos ofrece el español es un tesoro que merece ser estudiado con todo el detalle que la tecnología y los medios actuales nos brindan. En este sentido, también queremos aprovechar el proyecto Aracne para hacer un llamamiento para que editoriales, organizaciones e instituciones dediquen sus esfuerzos a la digitalización de sus fondos documentales y los pongan a disposición de investigadores y público general. Nadie entendería que el cuadro de *Las hilanderas* estuviera

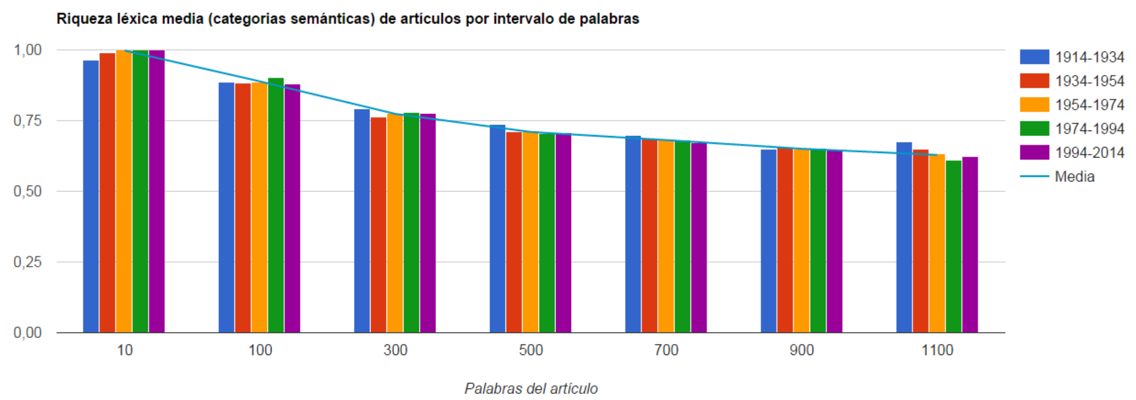
guardado bajo llave en un sótano y lejos del público; muy al contrario, la sociedad invierte un esfuerzo notable en cuidar, difundir y explotar un patrimonio que es de todos y cuya conservación y disfrute enriquecen a la sociedad en su conjunto. De igual modo, las hemerotecas son un tesoro que nos ofrece una información impagable sobre los más diversos aspectos de nuestra historia, de la evolución del país, del pensamiento, de las inquietudes como sociedad y, en definitiva, de la vida humana en colectividad. Historiadores, lingüistas, sociólogos, periodistas, documentalistas y especialistas e investigadores de muy distintas disciplinas se beneficiarían enormemente de disponer de este valiosísimo material, al que hoy no siempre es fácil acceder ni procesar.

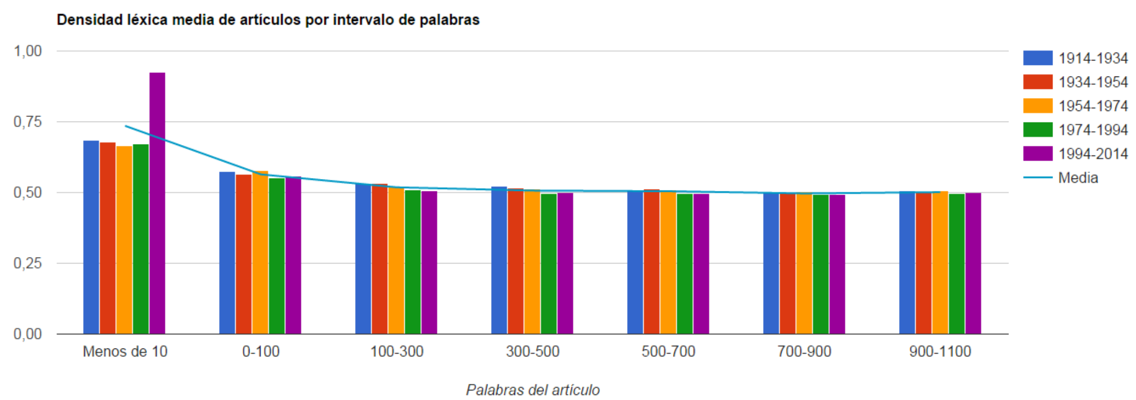
Con el proyecto Aracne, por tanto, damos por inaugurada una nueva línea de trabajo complementaria a la labor que ya realizamos en Fundéu BBVA: la del análisis científico, empírico y riguroso del uso del español en los medios de comunicación.

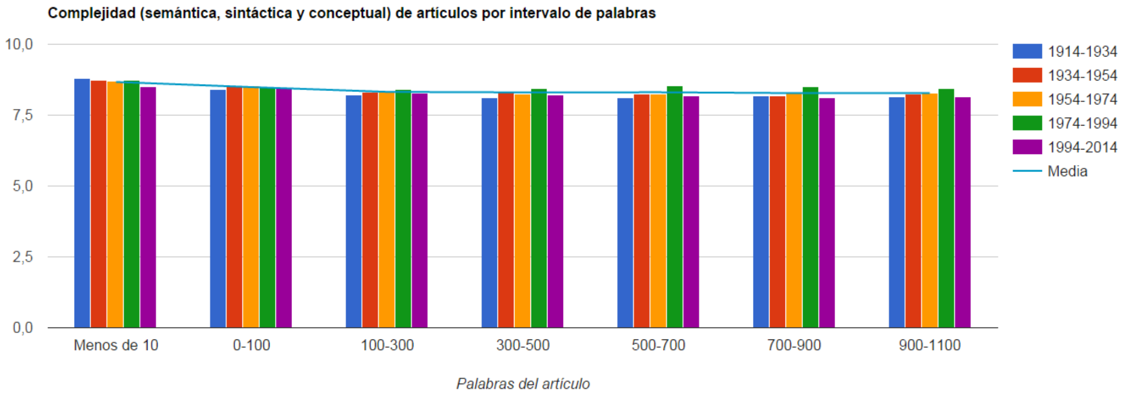
Anexo: gráficas

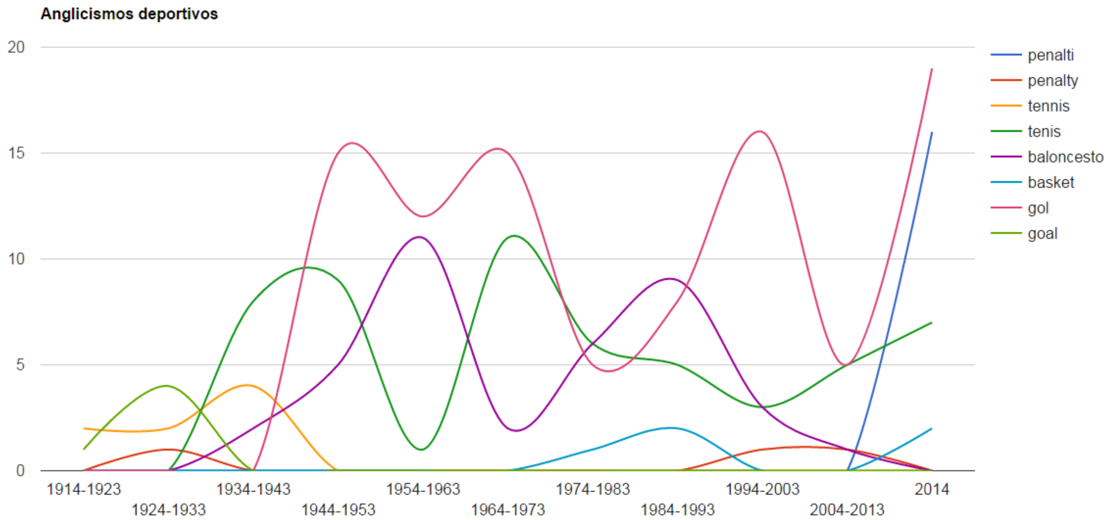
La información completa relativa a las gráficas está disponible en la sección del proyecto de la web de Fundéu BBVA. Los valores de frecuencia relativa están expresados por cada 100 000 palabras.

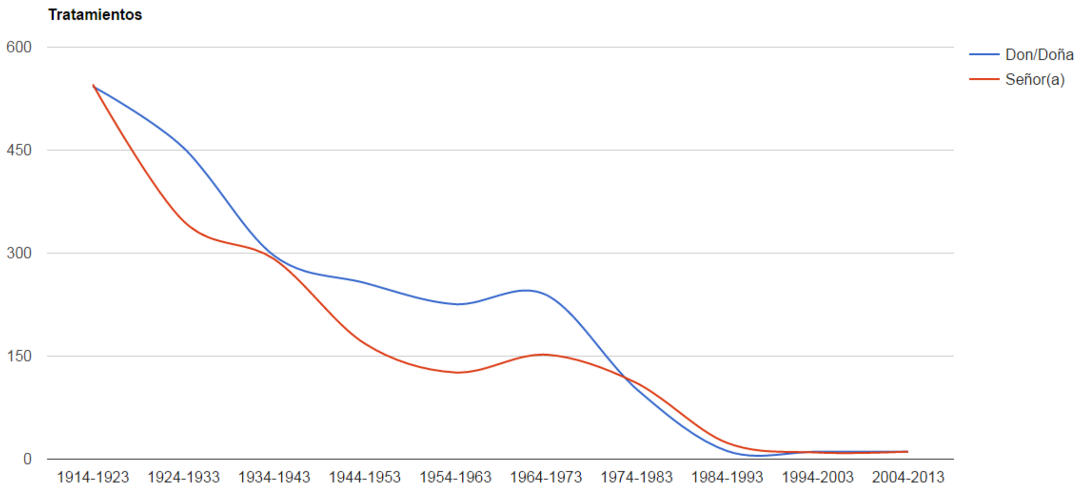


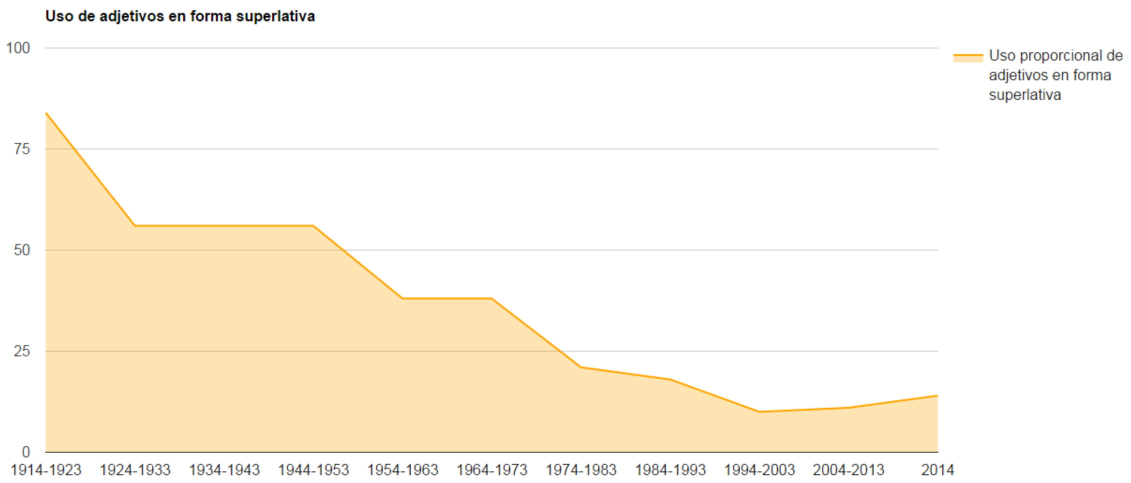


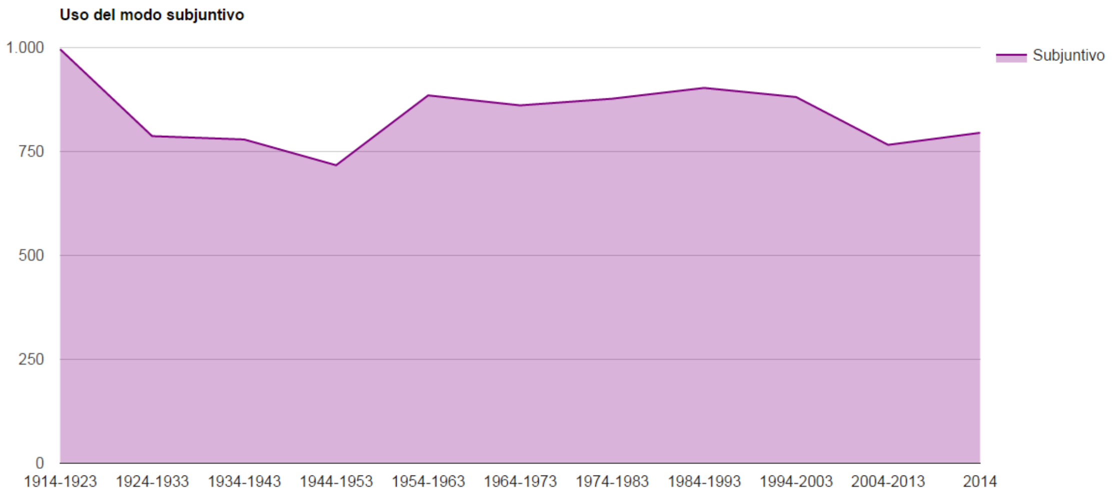


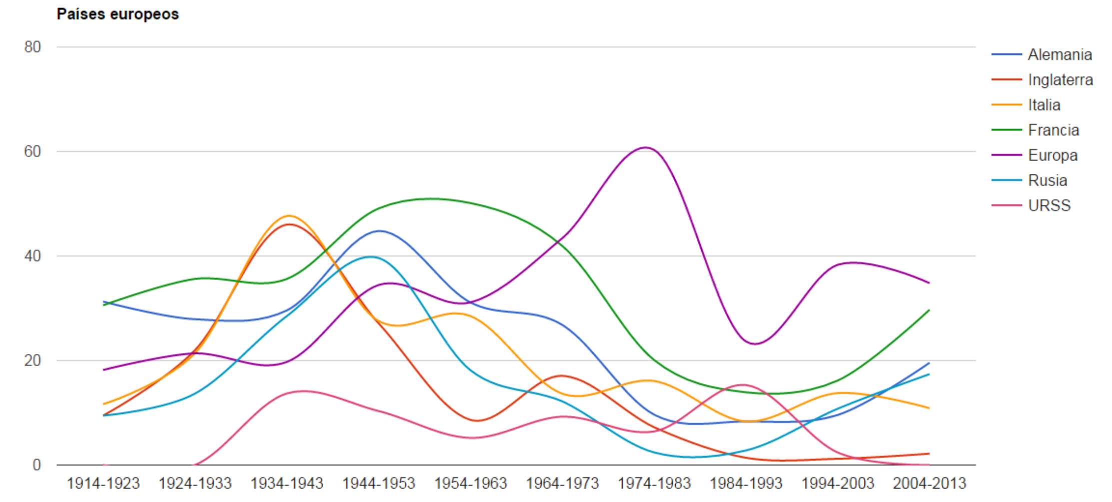

Complejidad (semántica, sintáctica y conceptual) de artículos por intervalo de palabras
10,0
7,5
5,0
2,5
0,0
Menos de 10
0-100
100-300
300-500
500-700
700-900
900-1100
Palabras del artículo
1914-1934
1934-1954
1954-1974
1974-1994
1994-2014
Media

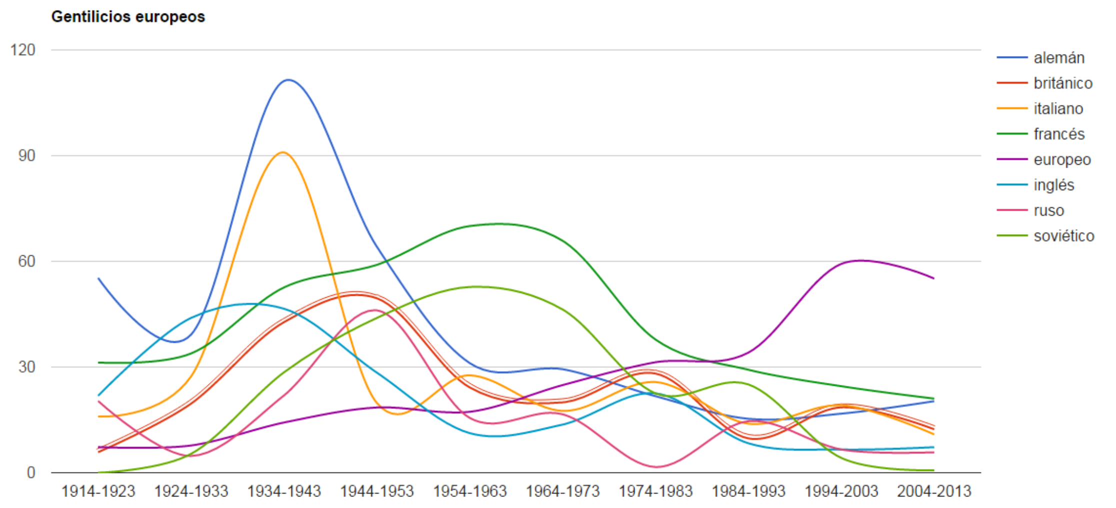

Anglicismos deportivos
20
15
10
5
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
2014
penalti
penalty
tennis
tenis
baloncesto
basket
gol
goal

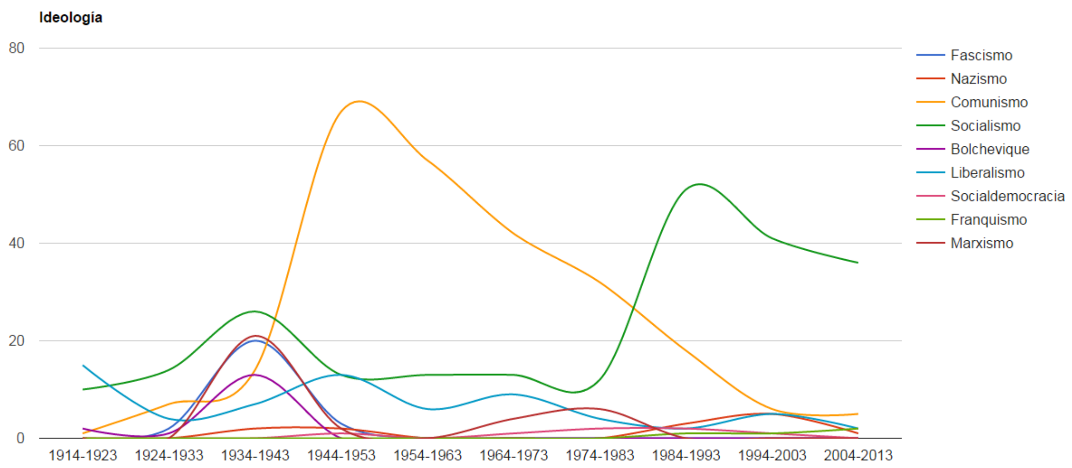

Tratamientos
600
450
300
150
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
Don/Doña
Señor(a)

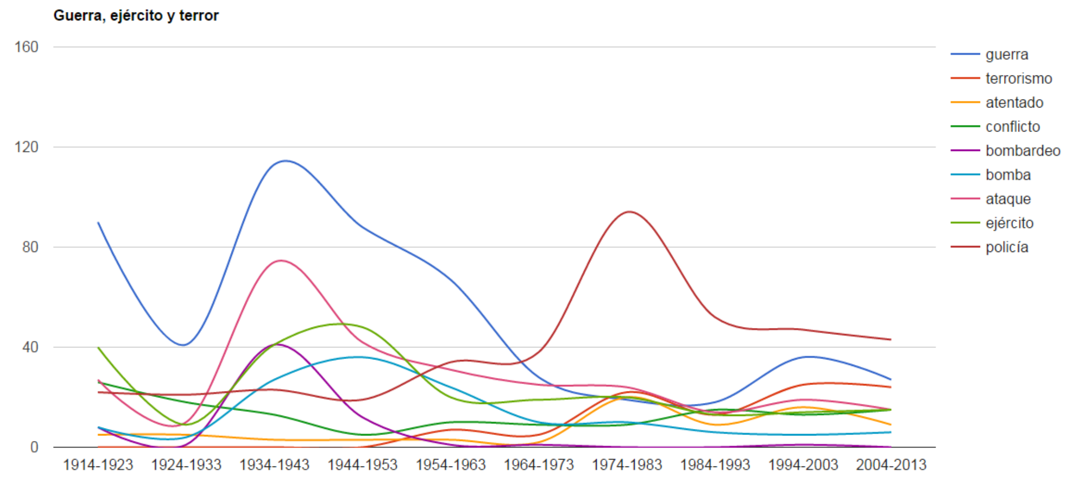

Uso de adjetivos en forma superlativa
100
75
50
25
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
2014
Uso proporcional de adjetivos en forma superlativa

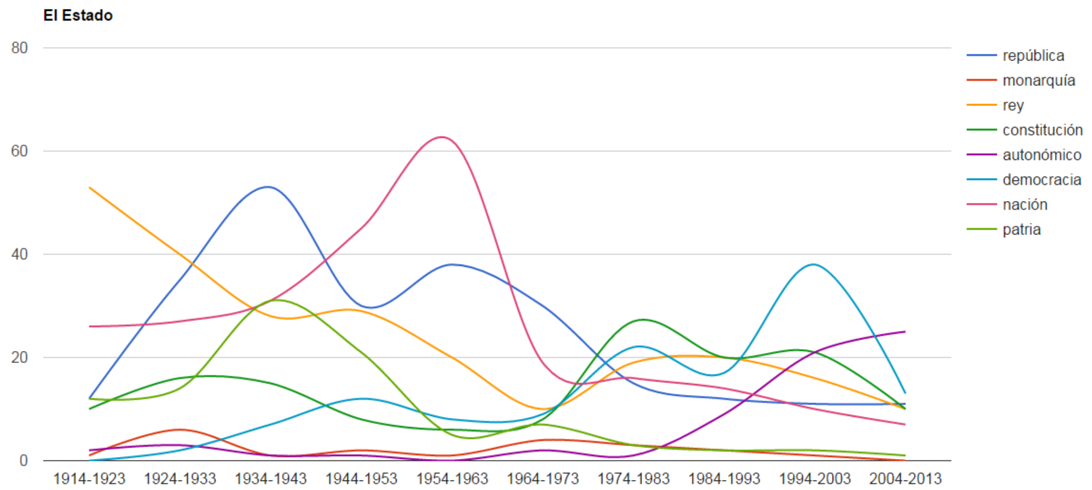

Uso del modo subjuntivo
1.000
750
500
250
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
2014
Subjuntivo

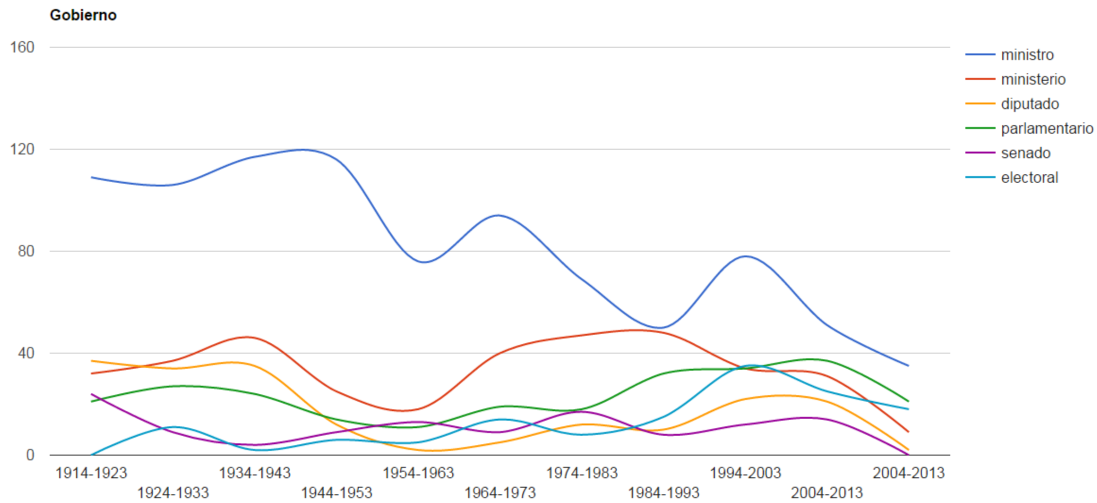

Países europeos
80
60
40
20
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
Alemania
Inglaterra
Italia
Francia
Europa
Rusia
URSS

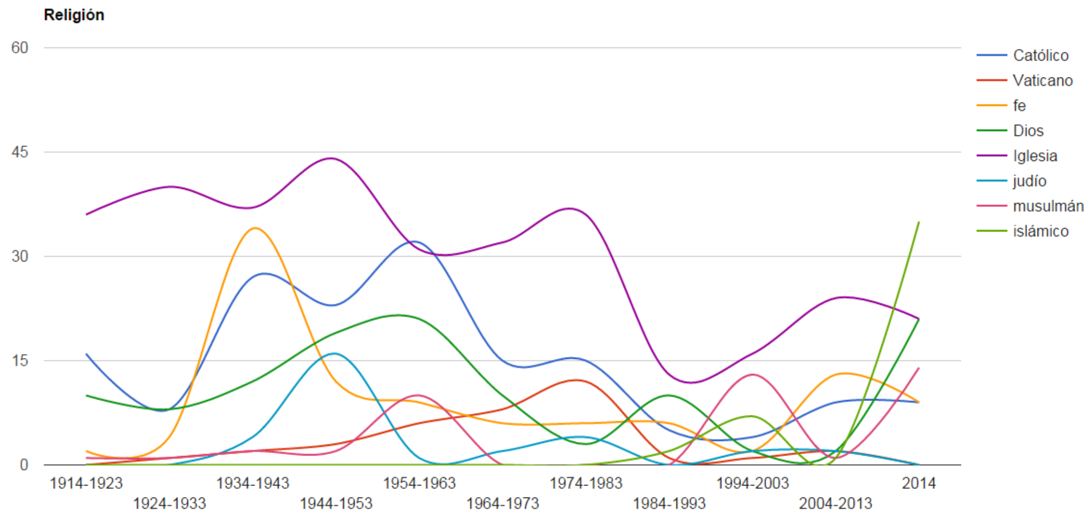

Gentilicios europeos
120
90
60
30
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
alemán
británico
italiano
francés
europeo
inglés
ruso
soviético

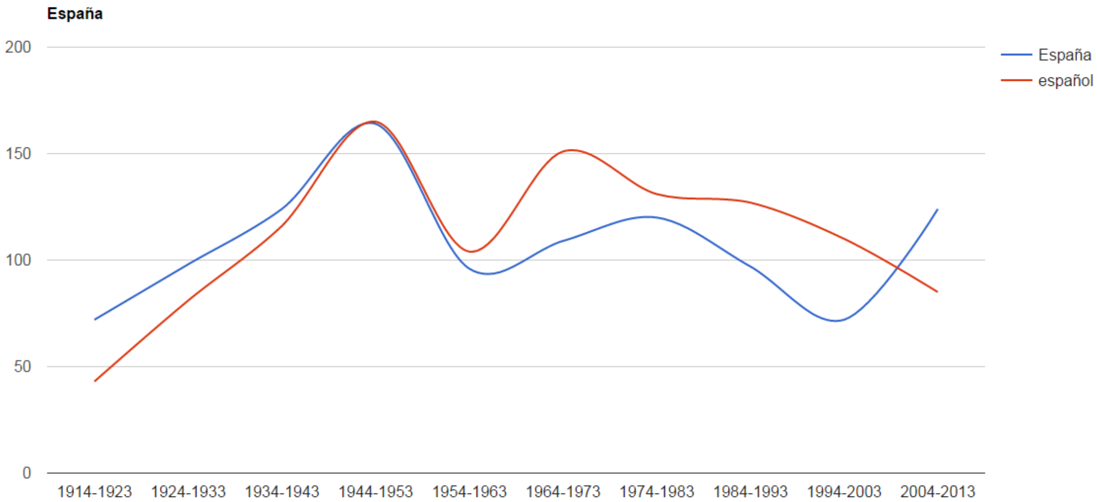

Ideología
80
60
40
20
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
Fascismo
Nazismo
Comunismo
Socialismo
Bolchevique
Liberalismo
Socialdemocracia
Franquismo
Marxismo

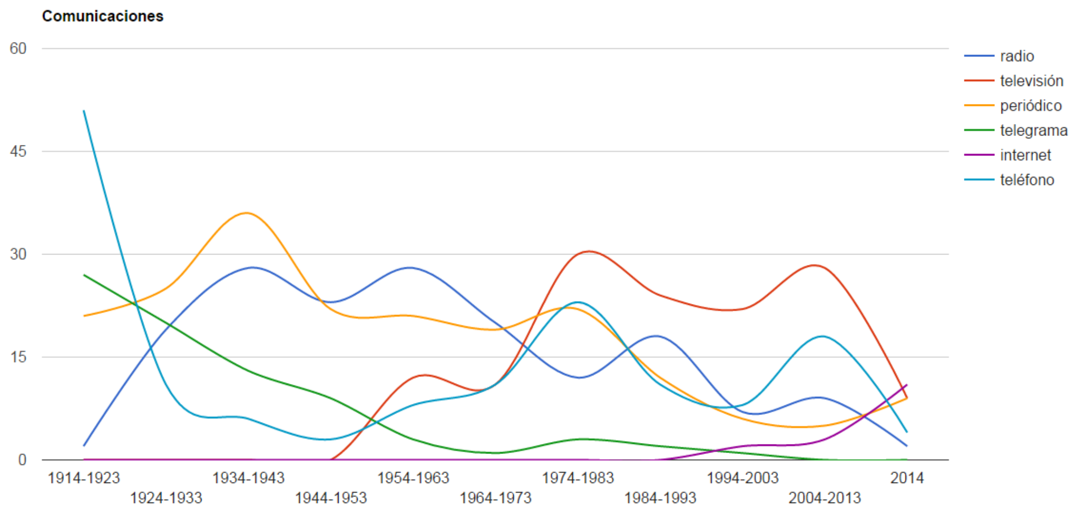

Guerra, ejército y terror
160
120
80
40
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
guerra
terrorismo
atentado
conflicto
bombardeo
bomba
ataque
ejército
policía

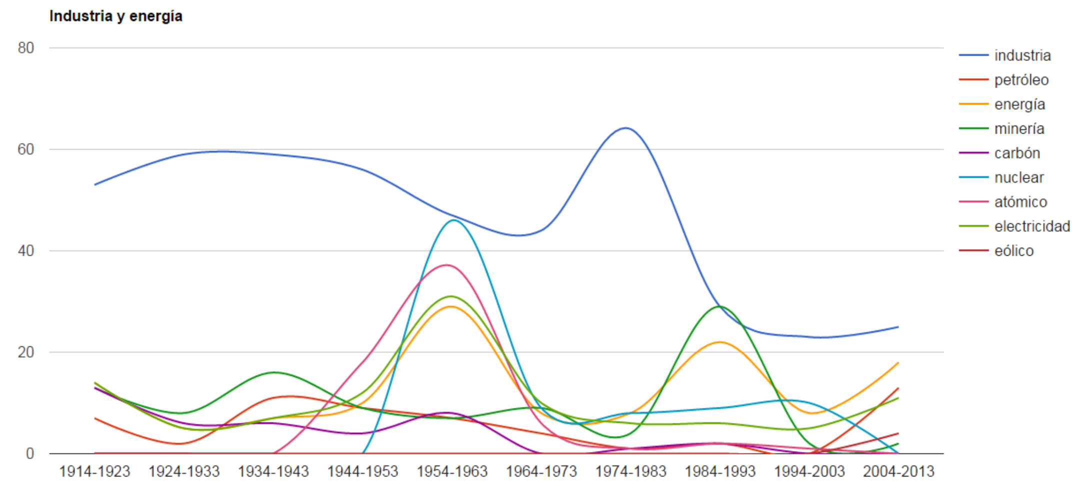

El Estado
80
60
40
20
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
república
monarquía
rey
constitución
autonómico
democracia
nación
patria

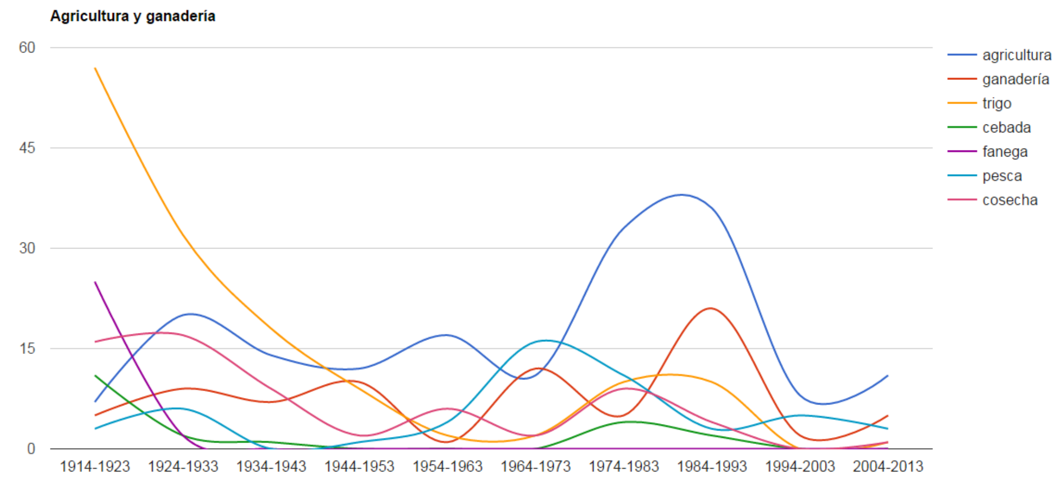

Gobierno
160
120
80
40
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
ministro
ministerio
diputado
parlamentario
senado
electoral

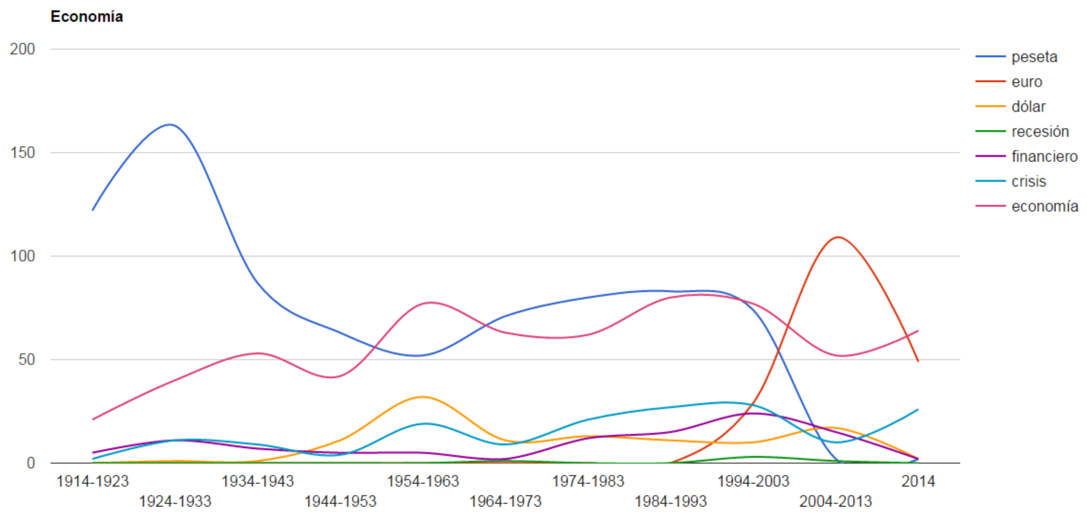

Religión
60
45
30
15
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
2014
Católico
Vaticano
fe
Dios
Iglesia
judío
musulmán
islámico

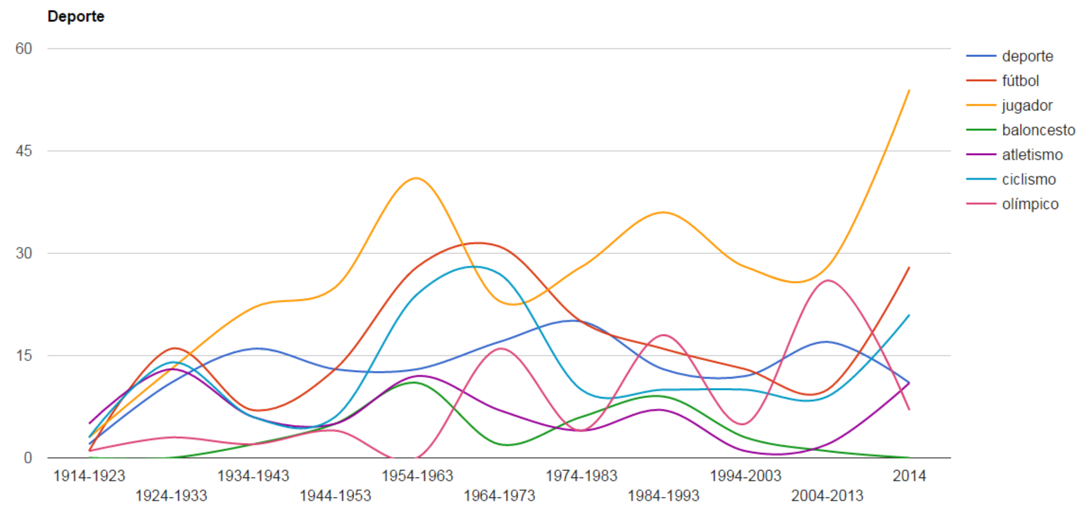

España
200
150
100
50
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
España
español

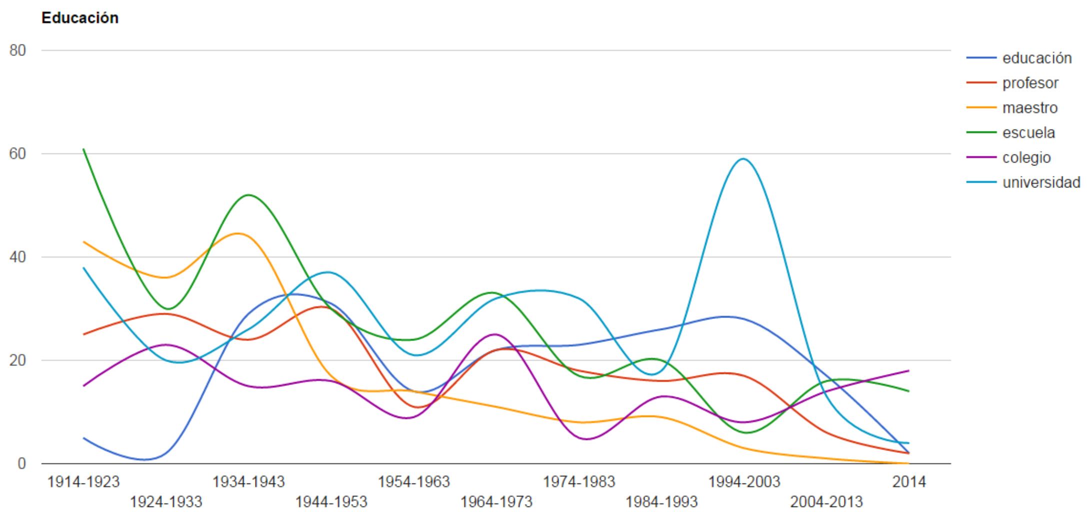

Comunicaciones
60
45
30
15
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
2014
radio
televisión
periódico
telegrama
internet
teléfono

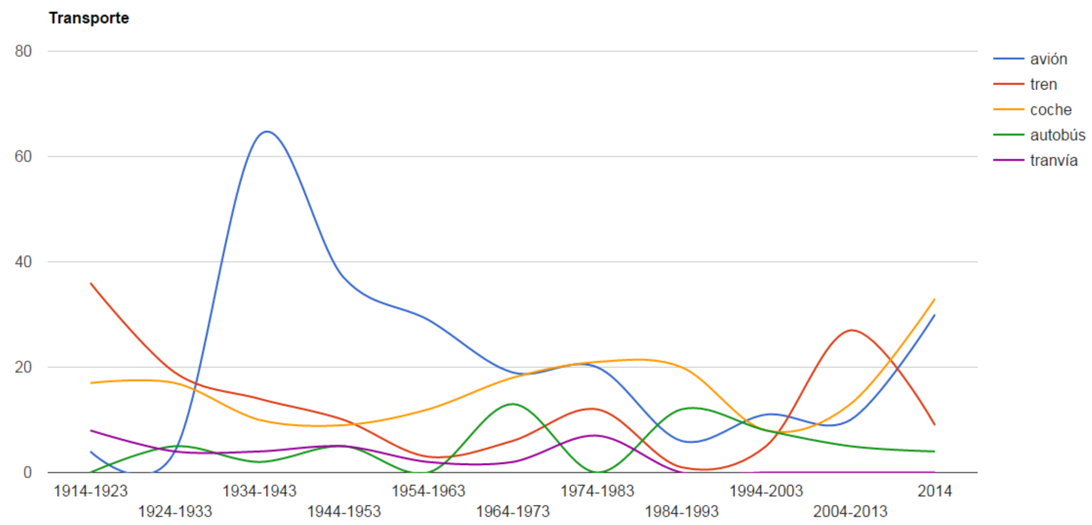

Industria y energía
80
60
40
20
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
industria
petróleo
energía
minería
carbón
nuclear
atómico
electricidad
eólico


Agricultura y ganadería
60
45
30
15
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
agricultura
ganadería
trigo
cebada
fanega
pesca
cosecha


Economía
200
150
100
50
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
2014
peseta
euro
dólar
recesión
financiero
crisis
economía


Deporte
60
45
30
15
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
2014
deporte
fútbol
jugador
baloncesto
atletismo
ciclismo
olímpico


Educación
80
60
40
20
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
2014
educación
profesor
maestro
escuela
colegio
universidad


Transporte
80
60
40
20
0
1914-1923
1924-1933
1934-1943
1944-1953
1954-1963
1964-1973
1974-1983
1984-1993
1994-2003
2004-2013
2014
avión
tren
coche
autobús
tranvía

Créditos, referencias y agradecimientos

Aracne es un proyecto de Fundéu BBVA, financiado por el banco BBVA y realizado con la tecnología lingüística de Molino de Ideas. El trabajo de documentación, corrección manual y supervisión de los textos ha sido llevado a cabo por Leticia Martín-Fuertes. El procesamiento, análisis de datos y dirección del proyecto lo ha realizado Elena Álvarez Mellado.

Queremos agradecer a *El Norte de Castilla, El Correo, Las Provincias,* al *Diario de Mallorca, Diario La Rioja, Heraldo de Aragón, ABC* y a la Biblioteca Nacional su colaboración desinteresada en Aracne. También queremos dar las gracias a Carlos J. Gil Bellosta por su asesoramiento y apoyo en los cálculos estadísticos del proyecto.

Bibliografía

Baayen, R. (2001). *Word frequency distributions*. Dordrecht: Kluwer Academic.

Baayen, R. (2008). *Analyzing linguistic data*. Cambridge, UK: Cambridge University Press.

Herrera-Soler, H., Martínez Arias, R. y Amengual, M. (2011). *Estadística aplicada a la investigación lingüística*. Madrid: EOS.

Johansson, V. (2008). *Lexical diversity and lexical density in speech and writing: a developmental perspective*. Lund University, Dept. of Linguistics and Phonetics, documento 53, pp. 61-79.

Malvern, D. (2004). *Lexical diversity and language development*. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.